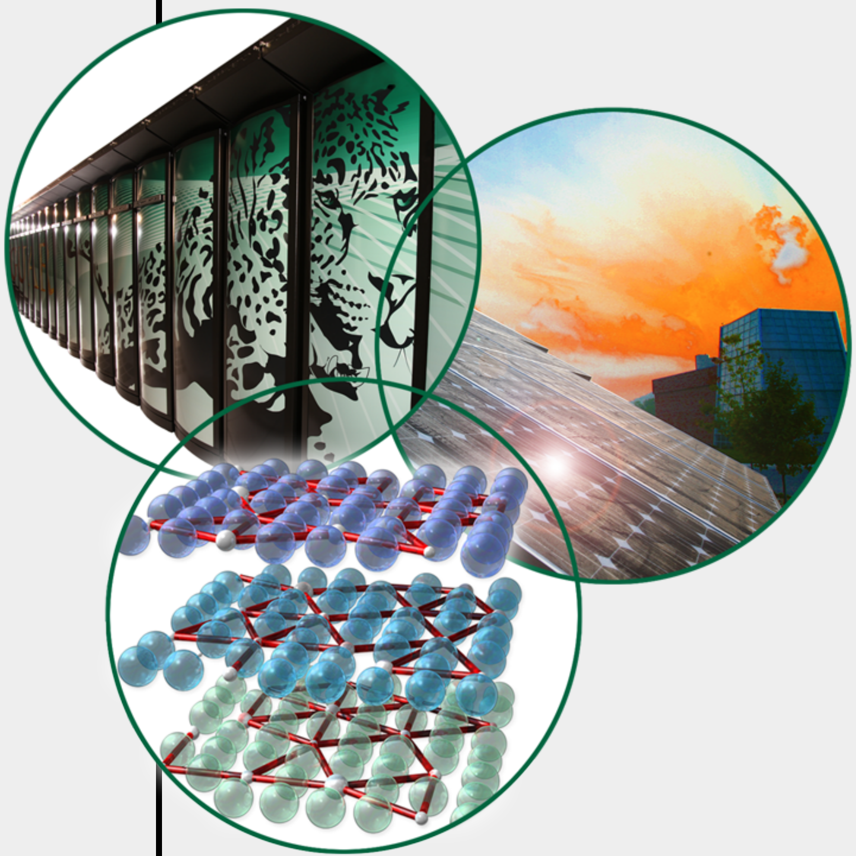
Metadata – Beyond Hierarchy and POSIX Attributes

Galen M. Shipman

HEC-FSIO Metadata Panel

8/8/2011

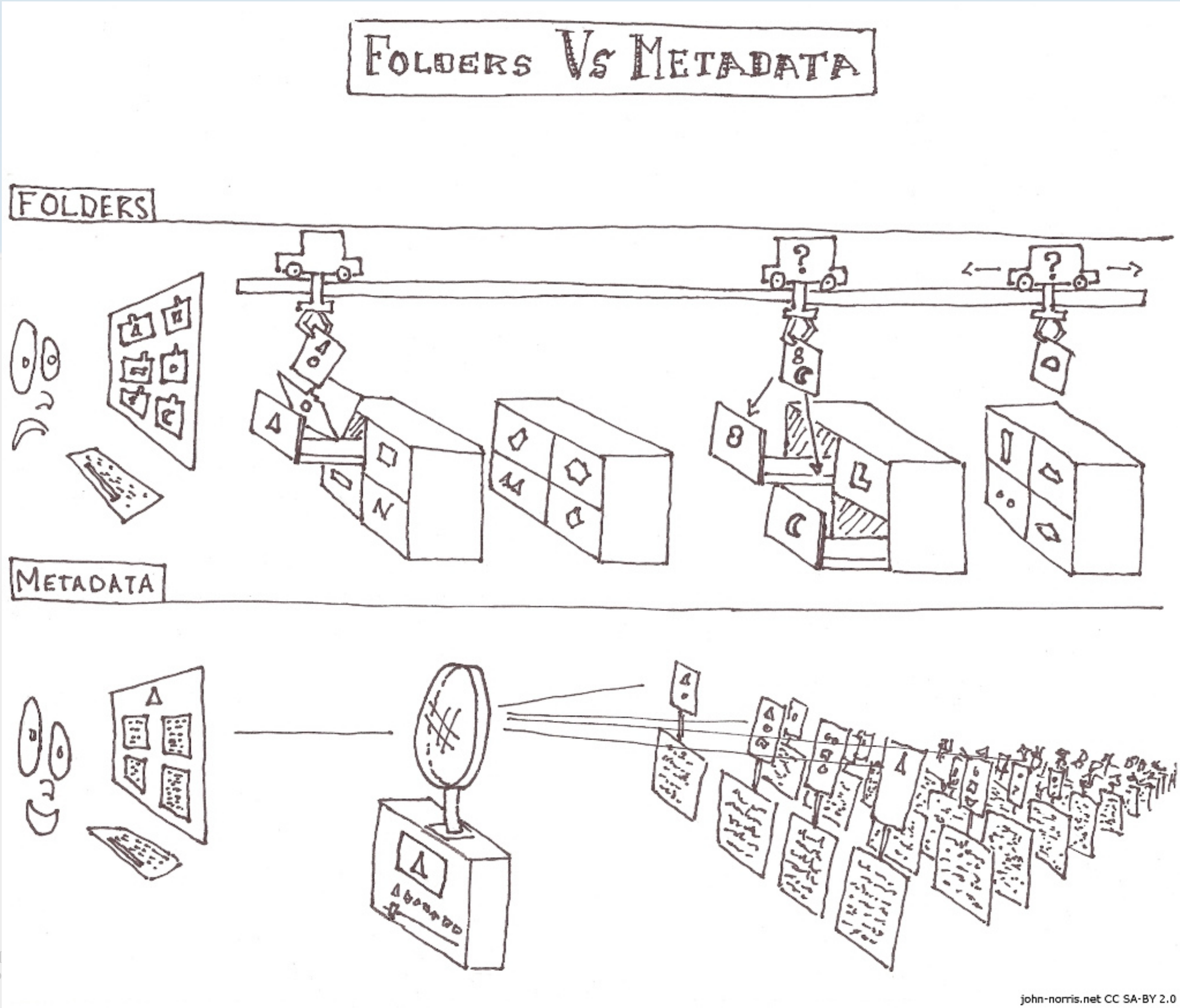
234281011 15384213 -rw-r--r-- 1 user foo\bar bam 0 2103687 "Aug 8 14:22:50 2011" "Aug 8 14:22:43 2011" "Aug 8 14:22:43 2011" "Aug 7 22:28:24 2011" 4096 4120 0 metadata.pptx



U.S. DEPARTMENT OF ENERGY

OAK RIDGE NATIONAL LABORATORY

MANAGED BY UT-BATTELLE FOR THE DEPARTMENT OF ENERGY


FOLDERS VS METADATA
FOLDERS
METADATA
john-norris.net CC SA-BY 2.0

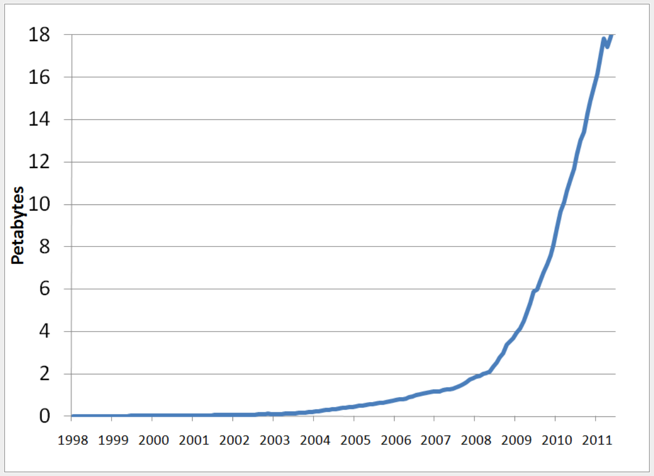
Metadata Explosion?

- How does the explosion of metadata influence the manner in which users might interact with storage systems?
  - Little influence, very little metadata is captured today in the Scientific Computing community through the use of sufficiently abstracted interfaces
- What can we learn from Spotlight and similar desktop tools?
  - Hierarchies are great in exploiting humans in the organization of data, humans are growing weary
  - Incorporating metadata harvesting as part of the I/O pipeline coupled with structured storage will free us from this exploitation
- Do you see an explosion of metadata compared to data size?
  - In the broader "Big Data" community, absolutely, representing relationships between granules often dwarfs data in size.
  - See first response for my answer in terms of Scientific Computing community
- Do you see an explosion of metadata dimensionality?
  - Only 6 dimensions exist (see POSIX attributes), to say otherwise is heresy

Managing the scientific data explosion

- Tens of thousands of disk drives
- Tens of thousands of tapes
- Over 25 Petabytes of data
- Over 200 million files
    - One user has over 400 TB of data in 8M files
    - One project has over 700 TB of data in 19M files
- Managed with very little information
    - User ID of owner
    - Group ID of owner
    - Total size in bytes
    - Time of last access ← current figure of merit!
    - Time of last modification
    - Time of last status change

Data growing exponentially



The POSIX Interface and Metadata

- A proven interface for human interaction
  - Hierarchical directories provide organization
  - Filenames provide a mechanism for identification
    - Augmented with standard attributes
  - But how often do you rely upon “spotlight” over “finder”?
  - Did you see Steve Poole’s desktop this morning?
- Widely used to support non-interactive “batch” workloads
  - We often see over 100 thousand files in single directory
  - Applications may use file naming strategies based on combinations of rank, timestep, variable identifier
  - Often very little information is conveyed in this organization and naming to a human

Structured data in an unstructured data store

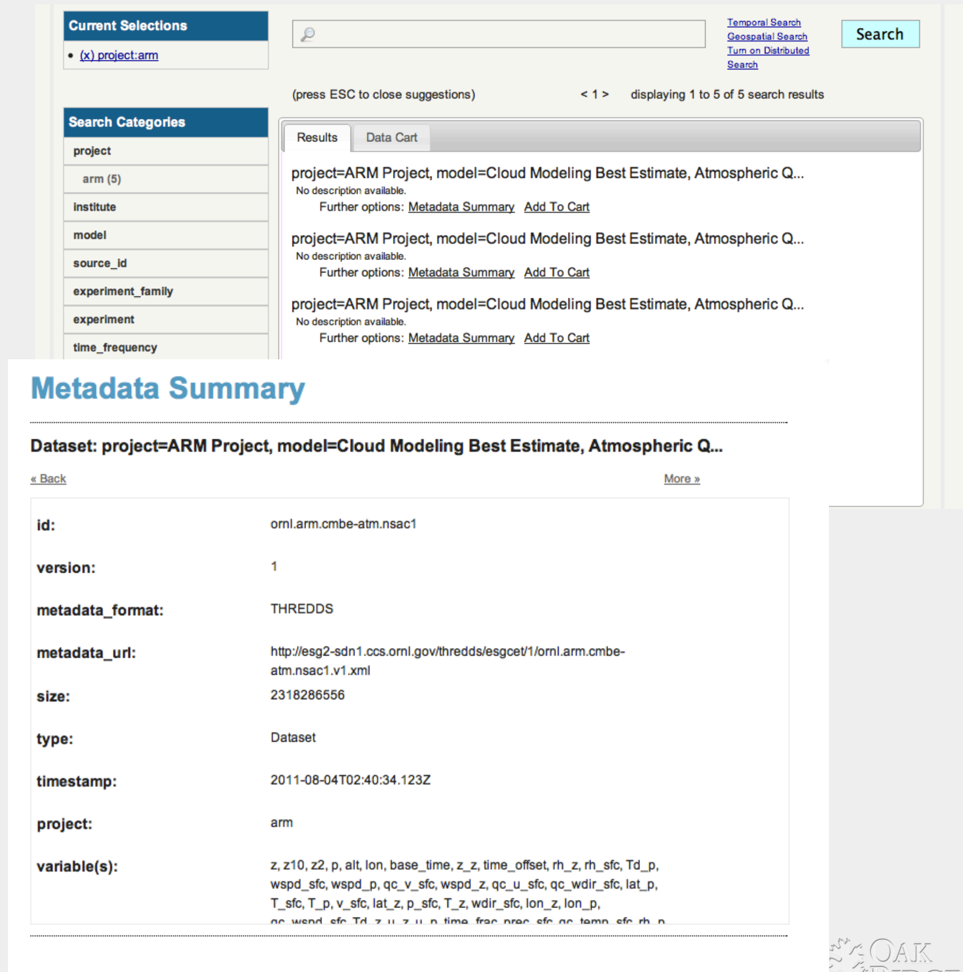
- The POSIX write/read/seek model is extremely flexible, supporting any number of data models
- This extreme flexibility often comes at the cost of understandability
- Scientific simulations often rely upon well known data models
  - But… this model is not imparted to the storage system
- Scientific datasets often have complex relationships that are not captured in scientific data models or storage systems
  - Climate land model experiment – land cover forcing – multiple scenarios
  - These datasets may comprise hundreds of thousands of files representing multiple model configurations with individual files spanning time and/or space

How do we impart meaning using file systems today?

- The climate community is an exemplar in data management for simulation data
- Data Reference Syntax (DRS) and Controlled Vocabularies
  - “atomic datasets” – granules mapped to individual variables representing the entire spatial-temporal domain
  - Variable names are defined by the Climate and Forecast Metadata convention
  - File names encode additional metadata:
    - filename = <variable name>_<MIP table>_<model>_<experiment>_<ensemble member>[_<temporal subset>].nc
  - Atomic datasets are then organized using directory structure
    - <activity>/<product>/<institute>/<model>/<experiment>/<frequency>/<modeling realm>/<variable name>/<ensemble member>/

How do we then share this information?

- Metadata from climate simulation datasets is then harvested into one or more THREDDS catalogs
- Search and discovery is enabled through Apache SOLR or Sesame RDF
- Data delivery is enabled through GridFTP or Data Mover Light



Lots of work to impart structure and meaning in an unstructured data store

- Can we impart structure and relations to better capture metadata directly within the data store? What is needed?
  - Need the ability to model complex relationships between data elements
  - Support for multi-dimensional data and metadata
  - Sparse data support
  - Flexible search capabilities
  - Distributed and parallel
- Exemplars exist: BigTable and Cassandra
  - In 2008 Google's largest BigTable instance contained 6 PB of data

Challenges

- Abandoning POSIX is painful but apparently rewarding
  - Why have data intensive industries been so successful in moving beyond POSIX?
- Current client interfaces are lacking (see Cassandra's Thrift)
  - Native Fortran, C, and C++ client interfaces would need to be built
- Messaging layer is not scalable in performance
  - Points to the need of a common communication API with high performance, scalability, and ubiquity
- Replication strategy is costly in bandwidth and space
  - Space likely to be less of a consideration as capacity gains outpace bandwidth improvements
  - Asynchronous replication during system idle times can reduce bandwidth requirements (write level one)